DARPA-BAA-14-21 Memex
Frequently Asked Questions (FAQs)
as of 4/2/14

87 Q. Does the 20-page proposal limit assume single-spaced text? Is there a required document margin size?

> A. The BAA does not specify spacing or margin requirements however, our preference is single-line spacing and one-inch margins.

86 Q. Referenced in paragraph 2, page 6 of the BAA, is the "common hardware" furnished by the Government?

> A. Yes. See Q/A 84.

85 Q. What is the difference between TA1 and TA2?

> A. TA1 is generally focused on discovery mechanisms, content processing, and indexing. TA2 is generally focused on query mechanisms, user interfaces, relevance rankings, and analytics. The line between these two technical areas is not tight, and teams are encouraged to submit the best possible proposal. Communicate your approach without getting too hung up on the division between technical areas.

84. Q. Who will host and who can access the source code repository?

> A. While the common repository will be managed/administered by one or more performers (selected based on the merits of their proposals), all performers will have access. DARPA will provide the hardware to support the common repository, so proposers do not need to include costs for this hardware.

83. Q. Does all source code need to be open source (can proprietary software be used)?

> A. We do not discourage proprietary software solutions. In order to meet the goals of the program, including wide adoption within the government and the ability to integrate with multiple contributors, liberal software licensing (e.g., Apache version 2) is encouraged; proprietary approaches must explain how they will be able to meet program goals. See Section VI.B.1 of the BAA.

82. Q. Do only combined TA1/2 proposals address integration component?

> A. All performers under the Memex program will be expected to work cooperatively with one another to develop, integrate, implement, test, and evaluate Memex capabilities. TA1+2 combined proposals will be responsible for building an integrated system across Technical Areas 1 and 2 to include the integration of other TA1 and TA2 components. Ultimately, all performers will contribute to the integrated system(s). One or several teams will have overall responsibility for the system(s).

81. Q: You mentioned the Internet often as the space should we look at. Are you also interested in sources outside of the Internet that may be freely available?

> A: The order of priority in which we're interested in data sources is: public data, application-based data, and then data which will best drive technology.

80. Q: In terms of the definition of domains—it seems that your program vision is to have a series of automated processes and then a very flat user space where the users are interacting with these automated processes in order to pull out domain-specific search. What are you seeing as the role of curators to find domains, to be intermediaries between that flat user space and the automated processes?

> A: That's an important role as the specification of the domain is an important problem. Discovery mechanisms are required to discover relevant content for the specified domain. Content discovery and information extraction is part of TA1.

79. Q: Why the specific exclusion regarding attribution?

> A: DARPA explicitly excluded this to avoid any misreading of the program's intention/goals.

78. Q: Will there be a client site or client-sponsored work location?

> A: The Government provided integration facility will serve as the client site.

77. Q: Will you consider existing commercial solutions or currently existing solutions?

> A: This BAA does not discourage the use of commercial or proprietary solutions. However, proposed research should investigate approaches that enable revolutionary advances in science, devices, or systems. Specifically excluded is research that primarily results in evolutionary improvements to the existing state of practice. Leveraging, enhancing, or refactoring existing open source software as part of an approach will be considered. Similar to proprietary software/technical data, proposers should clearly

explain why an existing software component will be used, as well as why the software was selected over alternative approaches. See Section VI.B.1 of the BAA.

76. Q: For TA2, access to domain experts and users will be essential. To what degree will TA3 provide access to them?

A: All performers will have access to the integration facility and this is where interaction with domain experts and users will take place.

75. Q: Are currently existing tech solutions precluded?

A: No. See Q/A 77.

74. Q: Are there any preferred requirements for the PI?

A: The BAA discusses anticipated areas of experience (e.g., cloud computing implementations and open platforms, etc.) as well as security clearance requirements (TA 3). Ultimately, the proposal should demonstrate how that particular PI, as well as the overall technical team, has the expertise and experience to accomplish the proposed tasks.

73. Q: What is the best way to package commercial proprietary technology in the proposals?

A: If proposers desire to use proprietary software or technical data or both as the basis of their proposed approach, in whole or in part, clearly identify such software/data in Volume I, Appendix A. See BAA Section IV.B.2.a.xii.(5) for submission details.

72. Q: In the case when commercial off the shelf tools are much more mature than the existing alternatives in the open source ecosystem, is it better to propose re-implementing that technology with a more open license, or is it better to license that proprietary technology so new higher level challenges can be addressed?

A: DARPA is interested in any solution to the problems outline within the Memex BAA, but it will not dictate one solution over another.

71. Q: Will DARPA provide datasets?

A: Per the BAA, DARPA intends to identify common, open public data sets to all performers for tests and evaluation.

70. Q: Is multilingual support required?

A: TA1 will be responsible for content discovery and information extraction relevant to the domain so proposers will need to determine whether multilingual support is required.

69. Q: Do TA2 performers have to select specific TA1 performers for their development of a query language?

> A: Collaboration among performers will be encouraged via a common working space, common furnished data, a common computational environment, and shared feedback from users. The query language and the specification of domain will be a joint exercise. You should propose what you think is best.

68. Q: Is the combined TA1+2 area solving a superset of TA1+TA2 problems, or just integration?

> A: It's both of those things. TA1+2 combined proposals will address each component as well as system integration.

67. Q: Given a platform of patented technology for a fully open and extensible knowledge base based on open standards, how important is it that the platform itself is open source?

> A: It depends on the merit of the proposal. If proposers desire to use proprietary software or technical data or both as the basis of their proposed approach, include a clear reason why and an explanation how the solution will fit within the program's goals, as well as any nonproprietary alternatives. The program will emphasize creating and leveraging open source technology and architecture. Intellectual property rights and software licenses asserted by proposers are strongly encouraged to be aligned with open source regimes.

66. Q: Do TA1/TA2 combined proposals have to provide an integration framework suitable for all, or just for the given approach?

> A: The TA1+2 combined proposers will need to provide an integration framework for all TA1 and TA2 components.

65. Q: Testing datasets?

> A: Per the BAA, DARPA intends to identify common, open public data sets to all performers for tests and evaluation.

64. Q: Can we submit a TA1 proposal and then find collaborators for TA2 and TA3?

> A: Yes, you may submit a TA1 proposal on your own and team with others on TA2 and/or TA3 proposals.

63. Q: Please let us know the acceptability of government purpose rights (GPR).

A: It is DoD policy to acquire only the rights to technical data and software that is necessary to satisfy the agency's needs. Per the BAA, proposers should follow all proposal submission instructions regarding the identification of any IP assertions that are less than unlimited rights. The acceptability of any IP assertions will be determined on a case-by-case basis per the information included within the proposal and how it conforms to the goals of the Memex program.

▲▲▲▲NEW FAQs▲▲▲▲

62. Q. Is there a preferred contract type (e.g., Cost Plus Fixed Fee)?

A. No.

61. Q. Is it only the Prime Contractor who is required to obtain a DUNS number, register in the System for Award Management (SAM), and register in E-Verify or do all Subcontractors, Vendors, and Consultants bid in the proposal have to complete these tasks prior to proposal submission?

A. Only the prime proposer needs to complete these tasks.

60. Q. If the Prime Contractor qualifies as an "Other Small Business," is the subcontractor plan then not applicable/not required?

A. In accordance with FAR 19.702 (b)(1), subcontracting plans are not required from small business concerns.

59. Q Have there been seedling projects or studies in preparation for this solicitation? If so, what were the topics, who executed them, and are reports available?

A. No additional information is available about preparatory efforts.

58. Q. Are research institutes and universities eligible to submit a proposal to Memex?

A. Yes.

57. Q. What approach should proposers take?

A. Proposers will need to decide on which approach they will take.

56. Q. In order to plan for travel of our various team members, could you please expand on the purpose of the quarterly and summer travel to DC, as opposed to PI meetings?

> A. Quarterly meetings will be a week long and address iterative development; more significant team development and integration will take place during the summer months. PI meetings are programmatic with updates to program management concerning plans, schedules, technical accomplishments, team, and funding status.

55. Q. Can foreign entities submit proposals to participate in the Memex research program?

> A. Yes.

54. Q. Do you plan to consider proposals that emphasize visualization and/or visual analytics systems?

> A. Yes; those are important technical components. User interface is part of TA2.

53. Q: Could you explain more about countering counter-crawling technology?

> A: We have to understand robots.txt and what is expected in terms of behaviors and terms of use at websites. We're trying to know those boundaries and approach them respectfully in an open way. We are interested in reorganizing public content.

52. Q: State sex offender registries all have stricter licensing that essentially says, "You're going to use this under pretty narrow terms." Indexing and caching data does not seem to be one of those narrow terms.

> A: The technology in TA1 and TA1+TA2 is about the data-agnostic application, the user-defined domain, and we need to be flexible in that space. On the application side, there will be measures taken to fully address the application that may involve relationships with other data providers that maintain Enterprise law enforcement databases or other government databases.

51. Q: The BAA talks about access to members-only forums. If it's an invitation-only type of thing, would you consider that publicly available?

> A: It would be open if the requesting party is invited. We're not going to break into forums; we're not hacking into systems.

50. Q: In other countries, there are different IP regimes that sometimes involve stricter characterizations of rights, especially of content within databases. How will you deal with that?

> A: The first answer is that we're operating within the U.S. under U.S. laws and jurisdictions. The international aspect of the effort is very important, especially on the application side, and we will adhere to all U.S. laws and regulations when it comes to the technology we're developing and the way it's used. The second answer is that Memex is focused on publicly available information from our vantage point, and the third is that it's an assessment. We need to understand in an automated way what the expected terms of use are, the robots.txt, and the intent behind those web spaces.

49. Q: You mentioned the Internet often as the space should we look at. Are you also interested in sources outside of the Internet that may be freely available?

> A: The order of priority in which we're interested in data sources is: public data, application-based data, and then data which will best drive technology.

48. Q: Clearly you want to have references such as company names and URLs but what are the sorts of attribution that you're not interested in?

> A: Per the BAA, the Memex program is specifically not interested in proposals for the following: attributing anonymous services, deanonymizing or attributing identity to servers or IP addresses, or gaining access to information which is not intended to be publicly available. For example, if we're indexing content that's behind Tor, and there are content pages that have IP addresses associated with them that have gone through onion routing, we're not interested in trying to trace back to the original space where they were posted.

47. Q: Are there domains that are outside the scope of TA1/TA2 (e.g., travel search, or product search, or shopping domains)?

> A: That should be the function of the proposal, but we're not interested in duplicating research that other projects are working on. The initial application domain is defined in the BAA and other domains will be considered.

46. Q: Is there any preference between purchasing hardware and buying time from other providers?

> A: Per the BAA, we will maintain a common cloud computing environment which is disconnected. While we expect proposers to budget for commercial web service cloud time as appropriate for their approach, large purchases of equipment are not encouraged.

45. Q: Is there a common repository for source code hosted in the environment?

> A: Yes. We'll have common hardware for testing. We'll also have common testing on an Amazon EC2-like cloud service.

44. Q: Are there transition partners already identified?

> A: Yes, there are many. While we are not identifying them at this time, proposers are welcome to recommend partners.

43. Q: Domain-specific extraction will require NLP. Is that in scope?

> A: Yes, but we will not invest in them as core research areas. There are other DARPA programs that are investing in text analysis, translation, image analysis, and we don't intend to duplicate those investments.

42. Q: Technology to assess user intent could involve both TA1 and 2. Is there a preference for where that technology is proposed?

> A: No. It depends on the merit of your proposal.

41. Q: How are you defining "Dark Web?"

> A: As used in the BAA, dark web content refers to things that are not indexed, behind hidden services, or not linked to an existing index.

40. Q: Are data analytic components required or desired for TA2?

> A: They are desired.

39. Q: What's the total budget, broken down per year by TA?

> A: It depends on the merit of the proposals.

38. Q: The BAA says the "Memex program is specifically not interested in proposals for . . . gaining access to information which is not intended to be publicly available" but also says that wanted capability includes "discovery of dark web content, hidden services, etc. Crawling should also be robust to automated counter-crawling measures, crawler bans based on robot behavior, human detection, paywalls, and members-only areas." These seem to be incompatible. Could you explain what is intended?

> A: We are not interested in hacking into systems. We are not interested in de-anonymizing systems or in obtaining and processing information which is not intended to be publicly available. It's a bright line. Deep web content and dark web content are

included within public content, including by discovery, invitation, or pay. Memex is about developmental research into open technology to establish a common tech base to do domain-specific indexing and domain-specific search.

37. Q: How do you anticipate the technology will be continually updated as the web changes?

A: We hope to develop a community around it by making it open source. We also intend to continue paying for it for the next three years. We think that if it’s useful and it gets adopted, other people will continue to pay for it, both on the government side as well as the commercial side.

36. Q: Will TA1 performers be required to demo their systems on a large-scale infrastructure, e.g. cloud-based?

A: Yes.

35. Q: Is TA1 focused on extraction and TA2 focused on processing?

A: I think that TA2 proposers need to understand how to process information as part of the workflow. If an extractor or a set of crawlers ends up in large databases with information, you need to understand how frequently they get updated, how large they are, how those factors dictate what latency constraints you have for interaction, how long you may need to do some things offline, or some things interactively. Workflow and process is important.

34. Q: Do you have examples of deep web crawlers or search tools?

A:Examples can be found on the Internet.

33. Q: Some DARPA programs emphasize technology stretch. Others focus on robust engineering. Where is Memex in this space?

A: For Technical Areas 1 and 2, the combined proposals and the core software, it’s closer toward reusable and robust software for user-specified domains. For some of the particular areas, like natural language processing, or for the application spaces, the research questions may be a bit further out. For domain-specification, it’s a relatively difficult question. It depends on what part of the problem you’re trying to address. But the end goal is reusable software that can address a user-specified domain.

32. Q: What breakthroughs, if achieved first, would produce the most value in combating human trafficking?

A: We think that science may be more effective than current methods for allocating resources for Health and Human Services, prioritizing law enforcement, and military use of resources both domestically and abroad. Understanding the way in which the Internet is used as a vehicle for labor and sex workers is a very important part of that decision-making process.

31. Q: Does DARPA plan to award contracts directly or use another organization as the agent?

A: We anticipate using an agent.

30. Q: If the government, in particular NIST, has evaluated things like in TREC, do you see this as being relevant to Memex?

A: Absolutely.

29. Q: This program is encouraging open source and publication. However, it is also targeting a specific adversarial application (human trafficking). Is there not a tension between these goals?

A: Yes, there is a tension there. That's why the Technical Area 3 groups need facility clearances and need to be able to have those sensitive discussions. We're not going to broadcast the details that involve the particular counter-trafficking application on the Internet, but our goal is to make the core technology that deals with crawling, extraction, and interface design open and available.

28. Q: Several aspects of indexing (TA1) involve user interfaces, which can border/interact with search (TA2). Do you have any hopes/desires for interfacing between TA1 and TA2?

A: Indexing technology tends to involve interface, the application of technology tends to involve interface, and there's an interface technical area. There are core code bases necessary to address all of these areas (TA1-3), and it's been our experience that those tend to be separate skill sets and academic communities.

27. Q: In the BAA, there's no mention of social networks like Twitter. Are they not of interest?

A: They are of interest, but there's a whole program of DARPA that does social media analysis. We're not going to invest heavily in an area where we're already investing.

26. Q: If company-university collaboration takes place, who is typically the lead, and who's the sub?

A:DARPA does not prescribe solutions, so there is no "typical" arrangement. Proposers have the flexibility to propose whatever teaming arrangement they decide best suits their approach.

25. Q: Is there a region of the world that we should focus on?

> A: The Internet.

24. Q: How much does the language of human traffickers vary from day-to-day commercial search language?

> A: We need to find out more specific answers. English tends to be the language of commerce, which involves sex trafficking. It may be different for labor trafficking, and it may be different depending on what region you're considering.

23. Q: Do proposers have to address all aspects of a Technical Area or can we address part of one?

> A: You can address part of one.

22. Q: Will a copy of the slides presented today be made available?

> A: No. The slides are taken directly from content in the BAA.

21. Q: Can you share the total program budget?

> A: No.

20. Q: Do you anticipate multiple awards on all TAs?

> A: That depends on the quality of the proposals.

19. Q: Can companies submit multiple proposals on one TA?

> A: Yes.

18. Q: Both TA2 and 3 appear to have a user interface component. Which TAs would you say have the primary UI responsibilities?

> A: TA2.

17. Q: The BAA mentions "As We May Think" and "Mother of all Demos," which cover a lot of ground. Can you comment on particularly relevant topics?

> A: The BAA is not asking people to propose things that were exactly in those documents or demos. They're mentioned for inspiration.

16. Q: Specifically what role do you envision for human users as part of the overall system architecture?

A: We envision them to be involved all the time: in the articulation of the domain, the strategy, the interactions, the interface design, the feedback, and the usage. Users are key, on the "hands on keyboards" side, on the program management and resource allocation side, on the tech side, and on the system design and deployment side.

15. Q: Are you interested in information organization approaches that link information across different applications?

A: Yes.

14. Q: Can international students participate?

A: Per the BAA, non-U.S. organizations and/or individuals may participate to the extent that such participants comply with any necessary nondisclosure agreements, security regulations, export control laws, and other governing statutes applicable under the circumstances.

13. Q: Will the experiments in the lab be run against the open web or some limited/controlled subset?

A: Both.

12. Q: Is change detection and monitoring of individual pages a priority? Or discovery of static data, possibly with an expanding and growing dataset? Which dynamics are of interest?

A: The polling, surveying, or sampling of web pages as part of an application depends on what the application is and the articulation of those analysis tasks. It depends on the application as well as on what the technology is trying to do.

11. Q: On the "Memex Premise" slide [from the Memex Proposers' Day], you use the term "links based on shared content." Can you explain what you mean by that?

A: Shared content is keywords or content of various kinds that are shared across webpages. Those could include proper names, company names, numbers, addresses, comments that are part of the source code of the webpage, mechanisms that the webpage has used to deploy content, like pop-ups, or other enabling media.

10. Q: Does cost-sharing increase the chance of being selected?

A: No.

9. Q: What is the percentage estimate for each Technical Area?

A: It depends on the quality of the proposals.

8. Q: Will an automated alert via email be sent when FAQs are updated?

A: No.

7. Q: What is the scale of the excised subset that you'd like to target?

A: It should be user-defined.

6. Q: Can you speak to the rates of ingest per day or per hour?

A: It relates to the domain. Depending on your analysis tasks, you may have a different required sampling rate.

5. Q: Is Memex about helping users make sense of their findings and create analytical products?

A: Yes.

4. Q: Is the preferred solution a web application versus a standalone application?

A: It depends on the proposal.

3. Q: Any security requirements for the front end?

A: Not initially, although for transition into other government use, things like authentication, PKI, CAC, etc. may be required.

2. Q: Should the solution work on mobile devices?

A: Not initially, but if there's a strong proposal we'll consider it.

1. Q: How do you plan to handle situations when a company may choose TA2 and team with a company submitting in TA1? Would there be a combo TA1/TA2 submission?

A: Each company can submit a separate TA1 or TA2 proposals or may submit a combined TA1/2 proposal. We are looking for the strongest proposals, so submit whichever arrangement you feel produces the strongest approach.